Theoretical Computer Science 28 (1984) 329–336
North-Holland

NOTE

COMPLETING BIPREFIX CODES

Dominique PERRIN
LITP, Université de Rouen, 76130 Mont Saint Aignan, France

Communicated by M. Nivat
Received July 1983
Revised September 1983

Abstract. A subset X of the free monoid A^* is called a biprefix code if it is both a prefix code and a suffix code. We prove in this note that any finite biprefix code is contained in a maximal biprefix code which is recognizable by a finite automaton.

1. Introduction

Several problems in automata theory are of the following kind: given an uncompletely specified automaton, belonging to some special family of automata, is it possible to complete it within this family? (see for instance [8]). The purpose of this note is to prove a result of the same kind: *any finite biprefix code is included in a maximal biprefix code recognizable (by a finite automaton)* (Theorem 3.2).

The same problem can be raised for other classes of subsets of a free monoid. It is of course trivial for prefix codes since any finite (resp. recognizable) prefix code is included in a finite (resp. recognizable) prefix code. The case of general codes is more difficult. It was proved by Restivo [6] that there exist finite codes that are not included in any finite maximal code. Restivo also conjectured that any finite code is included in a maximal code which is recognizable (by a finite automaton). This conjecture has just been proved by Ehrenfeucht and Rozenberg [3] who proved that, more generally, any recognizable code is included in a maximal recognizable code.

The proof of the result presented here heavily rests on the conjunction of two facts: first, Cesari [2] has deeply investigated the structure of finite maximal biprefix codes, extending the results of Schützenberger [9] who first studied these objects and of the present author [5]; it happens that his results extend to the recognizable case and lead very close to the solution of our problem. Second, the systematic use of formal power series in this matter proved to be a very powerful tool. This appeared to Berstel and the present author during the elaboration of a chapter on biprefix codes of a book on codes that is being prepared [1].

2. Biprefix codes

Let A be an alphabet, A^* the free monoid over A. We use the notation $A^+ = A^* - 1$, where 1 denotes the empty word.

We denote by $\mathbb{Z}\langle\langle A\rangle\rangle$ the ring of series with coefficients in $\mathbb{Z}$ and noncommuting variables in A. For a series S and a word $w \in A^*$, we denote by (S, w) the coefficient of w in S. For a subset X of A^*, we denote by $\underline{X}$ its characteristic series. It is the element of $\mathbb{Z}\langle\langle A\rangle\rangle$ defined by $(\underline{X}, x) = 1$ if $x \in X$ and 0 otherwise (for an introduction to series, see [4] or [7]).

A subset X of A^+ is a *prefix code* if $XA^+ \cap X = \emptyset$. It is a *suffix code* if $A^+X \cap X = \emptyset$. It is a *biprefix code* if it is both a prefix code and a suffix code.

Let $X \subset A^+$ be a biprefix code. We define the indicator I_X of X as the series

$$I_X = \underline{A}^*(1 - \underline{X})\underline{A}^* \tag{2.1}$$

Let

$$V_X = A^* - A^*X, \qquad U_X = A^* - XA^*. \tag{2.2}$$

An *X-interpretation* of a word $w \in A^*$ is a triple (v, x, u) such that $w = vxu$ with $v \in V_X$, $x \in X^*$, $u \in U_X$.

The following result shows that the indicator of a biprefix code has nonnegative coefficients.

Proposition 2.1. *Let $X \subset A^+$ be a biprefix code. The coefficient (I_X, w) of a word $w \in A^*$ in the indicator of X is equal to the number of interpretations of w.*

Proof. Since X is a prefix code, we have $\underline{XA}^* = \underline{X}\underline{A}^*$. Hence $\underline{U}_X = (1 - \underline{X})\underline{A}^*$. And symmetrically, since X is a suffix code $\underline{V}_X = \underline{A}^*(1 - \underline{X})$. We have therefore

$$I_X = \underline{V}_X\underline{A}^*, \tag{2.3}$$

$$I_X = \underline{A}^*\underline{U}_X. \tag{2.4}$$

Now, from $\underline{U}_X = (1 - \underline{X})\underline{A}^*$ we deduce that $\underline{A}^* = \underline{X}^*\underline{U}_X$. Substituting in (2.3) we obtain

$$I_X = \underline{V}_X\underline{X}^*\underline{U}_X. \tag{2.5}$$

The coefficients of $w \in A^*$ in the product $\underline{V}_X\underline{X}^*\underline{U}_X$ is precisely equal to the number of interpretations of w and the result is proved. □

Note that we have for all $u, v \in A^*$, the inequality

$$(I_X, uv) \geq (I_X, u) \tag{2.6}$$

since by (2.3), $(I_X, uv) = (I_X, uv)$ and, by (2.4), $(I_X, uv) \geq (I_X, v)$. For a subset X of A^* we define

$$H_X = \{u \in A^* \mid A^+u \cap X \neq \emptyset\}. \tag{2.7}$$

The set X is said to be *thin* if $H_X \neq A^*$. Any finite set is obviously thin and it can be shown that any recognizable code is thin (cf. [1]).

Theorem 2.2. *For any biprefix code $X \subset A^+$, the following conditions are equivalent:*

(i) *X is a thin maximal biprefix code.*

(ii) *The coefficients of I_X are bounded.*

Moreover, if $d_X = \max\{(I_X, w) \mid w \in A^*\}$ we have

$$H_X = \{v \in A^* \mid (I_X, v) \leq d_X - 1\}. \tag{2.8}$$

Proof. (i)$\Rightarrow$(ii). Let w be such that $X \cap A^+ wA^+ = \emptyset$. Suppose that the coefficients of I_X are not bounded; let $u \in A^*$ be such that $(I_X, u) \geq |w| + 2$. Then, by (2.6),

$$(I_X, wu) \geq |w| + 2.$$

Since (I_X, wu) is equal to the number of left factors of wu in $V_X = A^* - A^*X$, there is at least one left factor u' of u such that $wu' \in V_X$. Then the word $z = wu'$ has no right factor in X and since $w \notin H_X$, z is not a right factor of any element of X. One can show symmetrically that there exists a word $t \in A^*$ which has no left factor in X and is not a left factor of any element of X. Then it is easy to see that $X \cup tz$ is a biprefix code, contradicting the hypothesis that X is a maximal biprefix code.

(ii)$\Rightarrow$(i). Let $d_X = \max\{(I_X, w) \mid w \in A^*\}$. Let $v \in A^*$ be such that $(I_X, v) = d_X$. Then, since $\underline{X}\underline{A}^* = \underline{A}^* + (\underline{A} - 1)I_X$, we have, for all $a \in A$ and $u \in A^*$,

$$(\underline{X}\underline{A}^*, auv) = 1 + (I_X, uv) - (I_X, auv). \tag{2.9}$$

By (2.6), (I_X, uv) and (I_X, auv) are at least equal to (I_X, v). Hence, $(I_X, uv) = (I_X, auv) = d_X$. Substituting in (2.9) we obtain $(\underline{X}\underline{A}^*, auv) = 1$ or equivalently $auv \in XA^*$. We have therefore proved that, for any $u \in A^+$, $uv \in XA^*$. This implies that $v \notin H_X$ and therefore that X is thin. It also implies that X is a maximal prefix code since for any $u \in A^+$, we have $uv \in XA^*$, showing that $X \cup u$ is not a prefix code. Therefore X is a thin maximal biprefix code.

Let us finally prove (2.8). If $v \in H_X$, there exist u, $w \in A^+$ such that $uvw \in X$. Then uv has one more right factor in U_X than v, namely uv itself. Therefore, $(I_X, v) \leq (I_X, uv) - 1 \leq d_X - 1$. Conversely, if $(I_X, v) \leq d_X - 1$, let u be such that $(I_X, u) = d_X$. Then $(I_X, uv) = d_X$ and thus there exists a right factor $u' \in A^+$ of u such that $uv \in U_X$. Since X is a maximal prefix code, there exists a $w' \in A^+$ such that $u'vw' \in X$. Hence $v \in H_X$. □

Let X be a thin maximal biprefix code. The integer

$$d_X = \max\{(I_X, w) \mid w \in A^*\} \tag{2.10}$$

is called the *degree* of X.

The set

$$K_X = X \cap H_X \tag{2.11}$$

is called the *kernel* of X.

Example 2.3. Let $A = \{a, b\}$ and $X = a \cup ba^*b$. Then X is a recognizable maximal biprefix code. We have $V_X = 1 \cup a^*b$ and therefore $I_X = \underline{A}^* + a^*b\underline{A}^*$, that is

$$I_X = a^* + 2a^*b\underline{A}^*, \qquad d_X = 2, \qquad K_X = a. \tag{2.12}$$

Proposition 2.4. *Let X be a thin maximal biprefix code, $d = d_X$ its degree and $K = K_X$ its kernel. Then*

$$I_X = \inf\{d\underline{A}^*, I_K\}. \tag{2.13}$$

Proof. By definition, it follows that

$$I_X = \underline{A}^*(1 - \underline{X})\underline{A}^*, \qquad I_K = \underline{A}^*(1 - \underline{K})\underline{A}^*.$$

Let $w \in H_X$. Then any factor of w which belongs to X also belongs to K. Therefore, $(\underline{A}^*\underline{X}\underline{A}^*, w) = (\underline{A}^*\underline{K}\underline{A}^*, w)$. This implies that $(I_X, w) = (I_K, w)$. Moreover, we have $(I_X, w) \leq d - 1$ by (2.8). Therefore $(I_X, w) = \inf\{d, (I_K, w)\}$.

Now, if $w \notin H_X$, then $(I_X, w) = d$ again by (2.8). But since $K \subset X$, we have $(\underline{A}^*\underline{K}\underline{A}^*, w) \leq (\underline{A}^*\underline{X}\underline{A}^*, w)$ whence $(I_X, w) \leq (I_K, w)$. Again we have $(I_X, w) = \inf\{d, (I_K, w)\}$. □

As a corollary to Proposition 2.4 we deduce the following result whose first part was proved by Césari [2] for finite maximal biprefix codes.

Theorem 2.5. *A thin maximal biprefix code X is uniquely specified by its degree d_X and its kernel K_X.*

Moreover, X is recognizable iff K_X is recognizable.

Proof. By Proposition 2.4, given $d = d_X$ and $K = K_X$ we can compute I_X and therefore X since (2.1) is equivalent to

$$1 - \underline{X} = (1 - \underline{A})I_X(1 - \underline{A}). \tag{2.14}$$

If X is recognizable, then H_X is recognizable and therefore also $K = X \cap H_X$. Conversely, if K is recognizable, then, by (2.5),

$$I_K = \underline{V}_K\underline{K}^*\underline{U}_K.$$

Since K is a recognizable subset of A^*, so are $U_K = A^* - KA^*$, K^* and $V_K = A^* - A^*K$. Therefore $\underline{V}_K$, $\underline{K}^*$ and $\underline{U}_K$ are $\mathbb{N}$-recognizable series, so that their product I_K. Now I_X, as given by (2.13), is again an $\mathbb{N}$-recognizable series (see [4, p. 154]). Then $\underline{V}_X = I_X(1 - \underline{A})$ is the difference of two bounded $\mathbb{N}$-recognizable

series. Therefore, it is $\mathbb{N}$-recognizable (see [4, p. 154]). Therefore, the set V_X is recognizable and so is X since $X = AV_X - V_X$. □

Example 2.3 (*continued*). Since $K_X = a$, we have

$$I_K = \underline{A}^*(1-a)\underline{A}^*.$$

Therefore, for each word $w \in A^*$,

$$(I_K, w) = 1 + |w|_b.$$

This gives $I_X = \inf\{2\underline{A}^*, I_K\} = a^* + 2a^*b\underline{A}^*$.

3. Completion of biprefix codes

For a biprefix code X, we define

$$m_X = \max\{(I_X, x) \mid x \in X\}, \tag{3.1}$$

which is an integer or ∞.

The following theorem characterizes the kernels of thin maximal biprefix codes.

Theorem 3.1. *Let $d \geq 1$ be an integer. A set $K \subset A^+$ is the kernel of a thin maximal biprefix code of degree d iff it satisfies the following conditions*:

(i) *K is a biprefix code which is not maximal.*

(ii) $m_K \leq d-1$.

Proof. *Necessity*. Let X be a thin maximal biprefix code. Let $x \in X$ be such that $(I_X, x) = m_X$. If, for $u, v \in A^+$ we have $uxv \in X$, then ux has one more right factor in U_X than x, namely ux itself. Therefore

$$(I_X, uxv) \geq (I_X, ux) \geq m_X + 1,$$

a contradiction. This shows that $x \notin H_X$. The set $K = K_X$ is thus strictly contained in X and K is not a maximal biprefix code. Now, for any $x \in K$, we have, by (2.8), $(I_X, x) \leq d_X - 1$. By (2.13) we have $(I_X, x) = (I_K, x)$. Therefore $(I_K, x) \leq d_X - 1$. We have thus proved that K satisfies conditions (i) and (ii).

Sufficiency. Let $I \in \mathbb{Z}\langle\langle A\rangle\rangle$ be the series

$$I = \inf\{d\underline{A}^*, I_K\}. \tag{3.2}$$

For any $a \in A$, $w \in A^*$ we have $0 \leq (I_K, aw) - (I_K, w) \leq 1$ since aw has at most one more right factor in U_K than w. Then we also have $0 \leq (I, aw) - (I, w) \leq 1$. This shows that the series $(1 - \underline{A})I$ is the characteristic series of a set $U \subset A^*$:

$$\underline{U} = (1 - \underline{A})I. \tag{3.3}$$

Let $a, b \in A$ and $w \in A^*$. We show that

$$(\underline{U}, aw) = 0 \Rightarrow (\underline{U}, awb) = 0. \tag{3.4}$$

Suppose first that $(I, wb) \geq d$. Then $(I_K, wb) \geq d$ and also $(I_K, awb) \geq d$. Therefore, $(I, wb) = (I, awb) = d$ and $(\underline{U}, awb) = 0$.

Suppose now that $(I, wb) \leq d-1$ or, equivalently, that $(I_K, wb) \leq d-1$. Then we also have $(I_K, w) \leq d-1$. By the hypothesis $(I, aw) - (I, w) = (\underline{U}, aw) = 0$. But since $(I_K, w) \leq d-1$, we have $(I_K, aw) \leq d$. Therefore $(I, w) = (I_K, w)$ and $(I, aw) = (I_K, aw)$. We obtain $(I_K, aw) = (I_K, w)$. This means that aw has no more right factors in U_K than w. Therefore, $aw \notin U_K$ or equivalently $\overset{\bullet}{a}w \in KA^*$. This implies that $(I_K, awb) = (I_K, wb)$ and, since $(I_K, wb) \leq d-1$, we obtain $(I, awb) = (I, wb)$ and $(U, awb) = 0$. This proves (3.4).

We deduce from (3.4) that the set U contains all the left factors of its elements. In fact, $1 \in U$ since $(\underline{U}, 1) = (I, 1) = (I_K, 1) = 1$. If $uv \in U$ with $u, v \in A^+$, let $uv = awb$ with $a, b \in A$, $w \in A^+$. Then by (3.4) we have $aw \in U$. An easy induction on $|v|$ proves that $u \in U$.

Let X be the set

$$X = UA - U. \tag{3.5}$$

Then X is a prefix code: if $x \in X$, let $x = ua$ with $u \in U$, $a \in A$. All the proper left factors of x are left factors of u. They belong to U and are therefore not in X. We can turn (3.5) into an equality between series, writing: $\underline{UA} + 1 = \underline{X} + \underline{U}$, or

$$1 - \underline{X} = \underline{U}(1 - A). \tag{3.6}$$

By definition of U given in (3.3) this leads to

$$1 - \underline{X} = (1 - \underline{A})I(1 - \underline{A}). \tag{3.7}$$

The proof that the series $(1 - A)I$ is the characteristic series of a set V containing all the right factors of its elements is symmetrical to the above, corresponding proof for U. Then $1 - \underline{X} = (1 - \underline{A})V$ shows that X is a suffix code. Therefore, X is a biprefix code. By (3.7), the indicator I_X of X is equal to I. Since, by definition, I is bounded, we obtain that X is a thin maximal biprefix code by Theorem 2.2. Since K is not maximal, its indicator I_K is not bounded (see Theorem 2.2). Therefore, there exists a word $w \in A^*$ such that $(I_K, w) \geq d$. Then $(I, w) = d$. Hence $\max\{(I, w) \mid w \in A^*\} = d$ and the degree of X is equal to d.

Let us finally show that K is the kernel of X. We obviously have $K \subset H_X$ since, for any $k \in K$, $(I, k) \leq (I_K, k) \leq m_K \leq d-1$. It is therefore enough to prove that $X \cap H_X = K \cap H_X$ or equivalently that, for any $w \in H_X$,

$$(\underline{X}, w) = (\underline{K}, w). \tag{3.8}$$

We prove (3.8) by recurrence on $|w|$. If $|w| = 0$, both sides of (3.8) are zero. Let now $w \in H_X - 1$.

Since $(I, w) \leq d-1$, we have $(I, w) = (I_K, w)$. Hence $(\underline{A}^*\underline{K}\underline{A}^*, w) = (\underline{A}^*\underline{X}\underline{A}^*, w)$. Since all the factors of w are in H_X, we have, by induction hypothesis, $(\underline{X}, s) = (\underline{K}, s)$ for every proper factor of w. Therefore, $(\underline{X}, w) = (\underline{K}, w)$. This concludes the proof of the theorem. □

We now derive the main result of this note.

Theorem 3.2. *Let $Y \subset A^+$ be a finite biprefix code and $d = m_Y + 1$.*
There exists a unique recognizable maximal biprefix code with degree d and kernel Y.

Proof. First, since Y is finite, $m_Y = \max\{(I_Y, y) \mid y \in Y\}$ is finite. By Theorem 3.1 there exists a thin maximal biprefix code of degree $d = m_Y + 1$ whose kernel is Y. By Theorem 2.5, X is unique and recognizable. □

Example 3.3. Let $A = \{a, b\}$ and $Y = \{a, bb\}$. We have

$$(I_Y, a) = 1, \qquad (I_Y, bb) = 2.$$

Therefore, $m_Y = 2$. The unique recognizable maximal biprefix code with degree 3 and kernel Y is

$$X = a \cup bb \cup ba^+b^+a^+b.$$

The indicator of X is

$$I_X = a^* + 2a^*b^*a^* + 3a^*b^+a^+ba^*.$$

The set Y is also included in the maximal biprefix code of degree 2 of Example 2.3:

$$X' = a \cup ba^*b.$$

One may observe that $X \cap X' = Y$. This is a general fact: X' is the so-called *derivative* of X, whose indicator is obtained by decreasing by 1 the value of the coefficients of I_X equal to d_X. The intersection of X with its derivative is the kernel of X (see [2, 1]). In our example we obtain:

$$I_{X'} = a^* + 2a^*b^+a^* + 2a^*b^+a^+bA^* = a^* + 2a^*bA^*$$

as in (2.12).

We conjecture a result which is more general than Theorem 3.2: any recognizable biprefix code is included in a maximal recognizable biprefix code. Theorem 3.1 cannot be used to prove this conjecture. In fact, for such a set as

$$Y = ba^*b$$

the coefficients of I_Y are not bounded on Y. This means that Y is not the kernel of any recognizable maximal biprefix code. However, this set may obviously be

completed by adding the singleton $\{a\}$. This suggests the possibility of solving the above conjecture by first adding a kernel to Y and then turning the resulting set into the kernel of a maximal recognizable code.

Finally, it is worth mentioning that not any thin biprefix code can be completed to a thin maximal biprefix code. In fact the algebraic language

$$Y = \{a^n b^n \mid n \geq 1\}$$

is a thin biprefix code. It is not included in any thin maximal biprefix code because the degree of this code could not be finite. As a matter of fact, Y is included in the restricted Dyck set D_2, defined by

$$D_2^* = \{w \in \{a, b\}^* \mid |w|_a = |w|_b\},$$

which is a maximal biprefix code which is not thin.

Acknowledgment

The author would like to thank J. Berstel with whom many results of this note were obtained.

References

[1] J. Berstel and D. Perrin, *The Theory of Codes* (Academic Press, New York) to appear.
[2] Y. Césari, Propriétés combinatoires des codes biprefixes, in: D. Perrin, ed., *Théorie des codes* (LITP, Paris, 1979) pp. 20–46.
[3] A. Ehrenfeucht and G. Rozenberg, Any regular code is included in a maximal regular code, to appear.
[4] S. Eilenberg, *Automata, Languages and Machines, Vol. A* (Academic Press, New York, 1974).
[5] D. Perrin, Codes asynchrones, *Bull. Soc. Math. de France* **105** (1977) 325–404.
[6] A. Restivo, On codes having no finite completions, in: S. Michaelson, ed., *Automata Languages and Programming* (Edinburgh University Press, 1976) pp. 38–44.
[7] A. Salomaa and M. Soittola, *Automata Theoretic Aspects of Formal Power Series* (Springer, Berlin, 1978).
[8] M.P. Schützenberger, A remark on incompletely specified automata, *Inform. Control* **8** (1965) 373–376.
[9] M.P. Schützenberger, On a special class of recurrent events, *Ann. Math. Stat.* **32** (1961) 1201–1213.